Data & Visual Analytics

Duen Horng (Polo) Chau

Georgia Tech

CSE6242 / CX4242
Jan 7, 2014
POLO CHAU

Assistant Professor, School of Computational Science & Engineering
Adjunct Assistant Professor, School of Interactive Computing
Georgia Tech

poloat@gatech.edu  www.cc.gatech.edu/~dchau
Office: Klaus 1324  404-385-7682
Google Scholar (h-index: 17)  YouTube videos

<table>
<thead>
<tr>
<th>POSITIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 2012 –</td>
</tr>
<tr>
<td>Assistant Professor</td>
</tr>
<tr>
<td>School of Computational Science &amp; Engineering, Georgia Tech</td>
</tr>
<tr>
<td>Dec 2012 –</td>
</tr>
<tr>
<td>Adjunct Assistant Professor</td>
</tr>
<tr>
<td>School of Interactive Computing, Georgia Tech</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EDUCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 2012</td>
</tr>
<tr>
<td>Ph.D. Machine Learning</td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
</tr>
<tr>
<td>Thesis: Data Mining Meets HCI: Making Sense of Large Graphs</td>
</tr>
<tr>
<td>Award: Carnegie Mellon SCS Dissertation Award, Honorable Mention</td>
</tr>
<tr>
<td>Committee: Christos Faloutsos, Jason Hong, Niki Kittur, Jiawei Han (UIUC)</td>
</tr>
</tbody>
</table>
Course Staff

Instructor
Duen Horng (Polo) Chau
Assistant Professor, CSE
Office hour: Thu 3-4pm, Klaus 1324

TA
Robert Pienta, PhD student, CSE

TA
Long Tran, PhD student, CS
I Work with Large Graphs
I Work with Large Graphs

= Large Network Data
Internet
50 Billion Web Pages
Facebook
800 Million Users

Modified from Marc_Smith, flickr
Citation Network
250 Million Articles
Many More

Twitter
Who-follows-whom (500 million users)

Amazon
Who-buys-what (120 million users)

AT&T Cellphone Network
Who-calls-whom (100 million users)

Protein-protein interactions
200 million possible interactions in human genome

## Large Graphs I Analyzed

<table>
<thead>
<tr>
<th>Graph</th>
<th>Nodes</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>YahooWeb</td>
<td>1.4 Billion</td>
<td>6 Billion</td>
</tr>
<tr>
<td>Symantec Machine-File Graph</td>
<td>1 Billion</td>
<td>37 Billion</td>
</tr>
<tr>
<td>Twitter</td>
<td>104 Million</td>
<td>3.7 Billion</td>
</tr>
<tr>
<td>Phone call network</td>
<td>30 Million</td>
<td>260 Million</td>
</tr>
</tbody>
</table>
7±2

Number of items an average human holds in working memory

George Miller, 1956
Data → Insights
How to do that?

**COMPUTATION** + **HUMAN INTUITION**
## How to do that?

<table>
<thead>
<tr>
<th><strong>COMPUTATION</strong></th>
<th><strong>INTERACTIVE VIS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td>Summarization, clustering, classification</td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td>&gt;Millions of nodes</td>
<td>Thousands of nodes</td>
</tr>
</tbody>
</table>

Both develop methods for making sense of network data
How to do that?

**COMPUTATION**

Automatic

Summarization, clustering, classification

>Millions of nodes

**INTERACTIVE VIS**
How to do that?

**COMPUTATION**

Automatic

Summarization, clustering, classification

>Millions of nodes

**INTERACTIVE VIS**
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td>Summarization,</td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td>clustering,</td>
<td>Thousands of nodes</td>
</tr>
<tr>
<td>classification</td>
<td></td>
</tr>
<tr>
<td>Interaction,</td>
<td></td>
</tr>
<tr>
<td>visualization</td>
<td></td>
</tr>
<tr>
<td>Thousands of nodes</td>
<td></td>
</tr>
</tbody>
</table>
How to do that?

<table>
<thead>
<tr>
<th>COMPUTATION</th>
<th>INTERACTIVE VIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic</td>
<td>User-driven; iterative</td>
</tr>
<tr>
<td>Summarization, clustering,</td>
<td>Interaction, visualization</td>
</tr>
<tr>
<td>classification</td>
<td>Thousands of nodes</td>
</tr>
</tbody>
</table>

>Millions of nodes

Thousands of nodes
How to do that?

**COMPUTATION**

- Automatic
- Summarization, clustering, classification

**INTERACTIVE VIS**

- User-driven; iterative
- Interaction, visualization
- Thousands of nodes
“Computers are incredibly fast, accurate, and stupid.

Human beings are incredibly slow, inaccurate, and brilliant.

Together they are powerful beyond imagination.”
“Essentially, all models are wrong, but some are useful”

George Box
Logistics

Course homepage: poloclub.gatech.edu/cse6242/

Discussion, Q&A, find teammates: Piazza (link on homepage, soon)

Submission: T-Square
Course Goals

• Learn **scalable visual** and **computation** techniques and tools, for typical data types

• Learn how to **combine** both kinds of methods (how they complement each other)

• Gain **practical** know-how

• Gain **breath** of knowledge
Course Expectation

• Overview of scalable visual and computation techniques and tools

• Gain knowledge & experience (useful for jobs, research)

• Experience with designing and developing an interactive analysis tool
Schedule

See course homepage

poloclub.gatech.edu/cse6242/
Grading

• 3-4 homework assignments (40%)
• End-to-end analysis
• Techniques (computation and vis)
• Hadoop (+ other “big data” tools)
• Group project (50%) -- 3 to 4 people
• Participation (10%) -- in class, and on Piazza